

本周周报（5.11-5.17）

刘昊南

本周工作

1. 协助鸿辉精简了气象论文的压缩域体绘制、球面体绘制以及混合绘制三个部分的内容
2. 对 VAUD 平台数据库的选型做了初步的调研，VAUD 平台的数据主要的数据类型为时空数据，同时需要满足大数据存储与查询的要求，目前有三种数据库的候选方案

1) PosgtreSQL

借助于 PostGis 扩展，PosgtreSQL 被广泛应用于地理时空数据的存储，在做基于 GIS 信息的查询的功能十分强大，提供了较多的功能。

优点：应用案例较多，已经有广泛的应用

缺点：传统的关系型数据库，提供的事务和约束关系等特性对于我们的数据来说并没有什么用，且具有传统关系型数据库共同的缺陷，就是不适用于大数据的存储，当数据量变大单台数据库服务器无法容纳时，需要手工的对数据库做分片和分区等复杂的操作来添加新的数据库服务器，且每次对数据库服务器集群扩容时都需要这样手工操作

2) MongoDB

MongoDB 实现了 GeoSpatial 的索引，能够对地理空间信息做一些简单的查询操作。

优点：MongoDB 是 document oriented 的 NoSQL 数据库，document 没有的 schema，不同的 document 也并没有约束或引用关系，这些正是我们的出租车数据、手机基站数据所具有的特点，它们之间本身并没有约束或引用关系，比较适合 NoSQL 数据库；同时，作为 NoSQL 数据库，设计之初就考虑到了大数据、高并发和分布式的应用场景，能够作为数据库集群来使用，数据库集群也能够容易地线性增长

缺点：MongoDB 的 GeoSpatial 索引上的查询功能较为简单，没有 PosgtreSQL 的查询功能丰富

3) GeoMesa

GeoMesa 实现了一种建立在 bigtable 数据库上的分布式的时空索引，支持海量的地理空间数据的插入、索引和查询，支持 Apache Accumulo、Apache HBase 和 Google BigTable 等 BigTable 衍生数据库

优点：GeoMesa 实现了一种索引，能够建立在 Apache Accumulo、Apache HBase 数据库上，这两种数据库都是建立在 Hadoop 的 HDFS 分布式文件系统上的，即我们可以直接在 Spark 集群本地的 HDFS 上建立数据库供 Spark 集群直接使用，而不需要另外再建立一个单独的数据库，通过网络传输来拉取数据

缺点：目前该项目的文档很少，参考资料稀缺

下周计划

1. 确定数据库选型、数据结构与数据组织的设计与实现